Explainable Face Recognition: Supplementary Material

Jonathan R. Williford[1][0000−0002−9178−2647], Brandon B. May[1][0000−0002−9914−2441], and Jeffrey Byrne[1,2][0000−0001−8973−0322]

[1] Systems & Technology Research, Woburn, MA 01801, USA
https://www.stresearch.com
{jonathan.williford,brandon.may}@stresearch.com
[2] Visym Labs, Cambridge, MA 02140, USA
jeff@visym.com

1 Supplementary Material

1.1 Qualitative Visualization Study

The inpainting game provides a quantitative comparison of XFR algorithms, however it does not provide insight as to how useful these XFR algorithms are on novel face images. In this section, we provide a qualitative study of XFR algorithms visualized on a standard set of triplets. We consider two target networks: ResNet-101 and Light-CNN [2], and provide visualizations for the whitebox methods referenced in the main submission. This analysis includes the following figures showing qualitative visualization results for combinations of (target network, XFR method): (ResNet-101, EBP, Fig. 3), (ResNet-101, cEBP, Fig. 4), (ResNet-101, tcEBP, Fig. 5), (ResNet-101, Subtree, Fig. 6), (Light-CNN, EBP, Fig. 8), (Light-CNN, cEBP, Fig. 9), (Light-CNN, tcEBP, Fig. 10), (Light-CNN, Subtree EBP, Fig. 11). Finally, we show results for the Light-CNN using only single probes (Fig. 12), or repeated probes (Fig. (13) to highlight the effect of non-mates in the triplet visualization.

From this visualization study, we draw the following conclusions:

1. **Non-localized.** Unlike facial examiners which leverage the complete FISWG standards for facial comparison, there is no evidence that modern face matchers leverage localized discriminating features such as scars, marks and blemishes. All visualizations are centered on the facial interior, and almost no activation is on the shape of the head. Also, the systems tend to overgeneralize to represent all faces in a standard manner using the eyes and nose, brow and mouth, ignoring localized features such as moles or facial markings.
2. **Pose variant.** The target networks tested are not truly pose invariant. When considering different probes of the same subject, where the probe differs in pose, the whitebox systems can generate different visualizations. This suggests that the underlying network is still pose variant.
3. **Triplet specific.** The features that are used for recognition depend on the selection of the triplet, notably the selection of the non-mate for comparison.

Fig. 1. Qualitative visualization study. This figure shows the XFR saliency maps generated using the LightCNN Subtree EBP method for 16 probes (columns) of 7 subjects (rows), each with 16 mates (not shown) and a common set of 8000 nonmates (not shown) all sampled from VGGFace2 [1]. Results show that the discriminative features used to distinguish a subject from the entire nonmate population are inconsistent, but are primarily the nose and mouth for frontal probes, including the eyes for non-frontal probes. See supplemental Fig. 15 for additional examples.

The visualized features are more consistent when considering a larger nonmate set (Fig. 1).

4. **Network specific.** The visualized features are dependent on the selected target network for visualization. A higher performing network (light-CNN) tends to use more facial features of the brow and mouth in addition to the eyes and nose, than a lower performing network (ResNet-50). No networks yet tested use the hair or chin.

References

1. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. 2, 16
2. X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 1, 9, 10, 11, 12, 13, 14

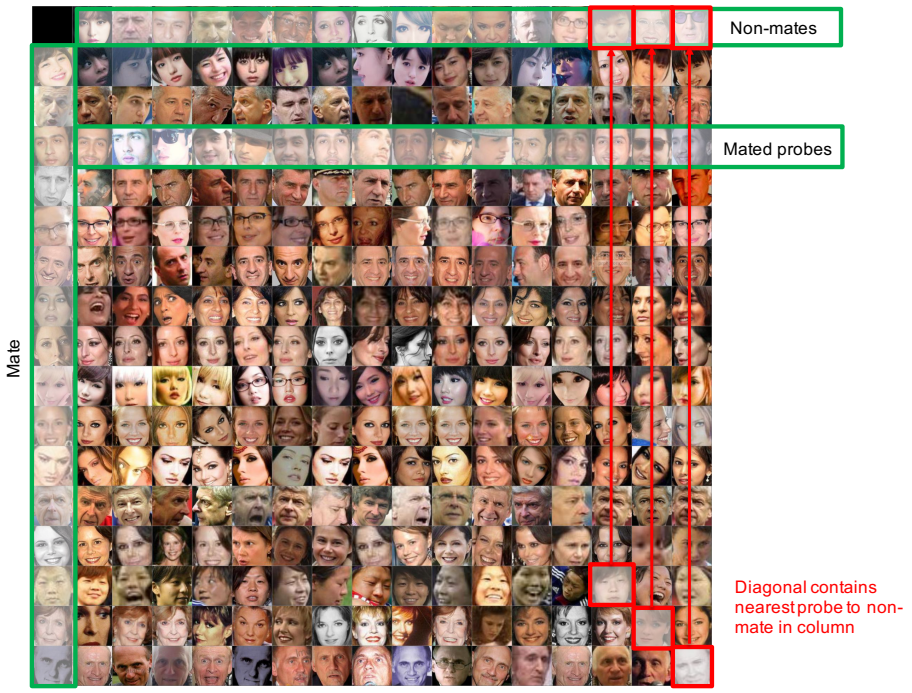

Fig. 2. Whitebox visualization overview. This montage shows a set of 16 randomly selected subjects from IJB-C, such that every row has the same identity. Images (i, j) in this montage define a triplet (m_i, p_{ij}, n_j) for probe p_{ij}, mate m_i in the first entry of column i and non-mate n_j in the first entry in row j. Non-mates are ordered such that on the diagonal are the nearest non-mated subject in IJB-C. In other words, for triplet (m_i, p_{ii}, n_i), non-mate n_i is more similar to m_i than any other nonmate n_j, using a ResNet-101 matching system. This montage is used to visualize how the whitebox saliency map changes when considering different triplets.

Fig. 3. EBP (ResNet-101). This montage is the same images as in Fig. 2, but with a whitebox saliency map derived from excitation backprop for a whitebox ResNet-101 system. Observe that EBP always selects the eyes and nose no matter what non-mated subject is being considered. This does not provide subtle distinctions between the regions that are discriminative for a mate vs. a non-mate, but it does provide a visualization of the regions of the probe that are used for classification. This visualization should be compared with Fig. 8 for the same subjects and whitebox method, but a different underlying trained network (light-cnn).

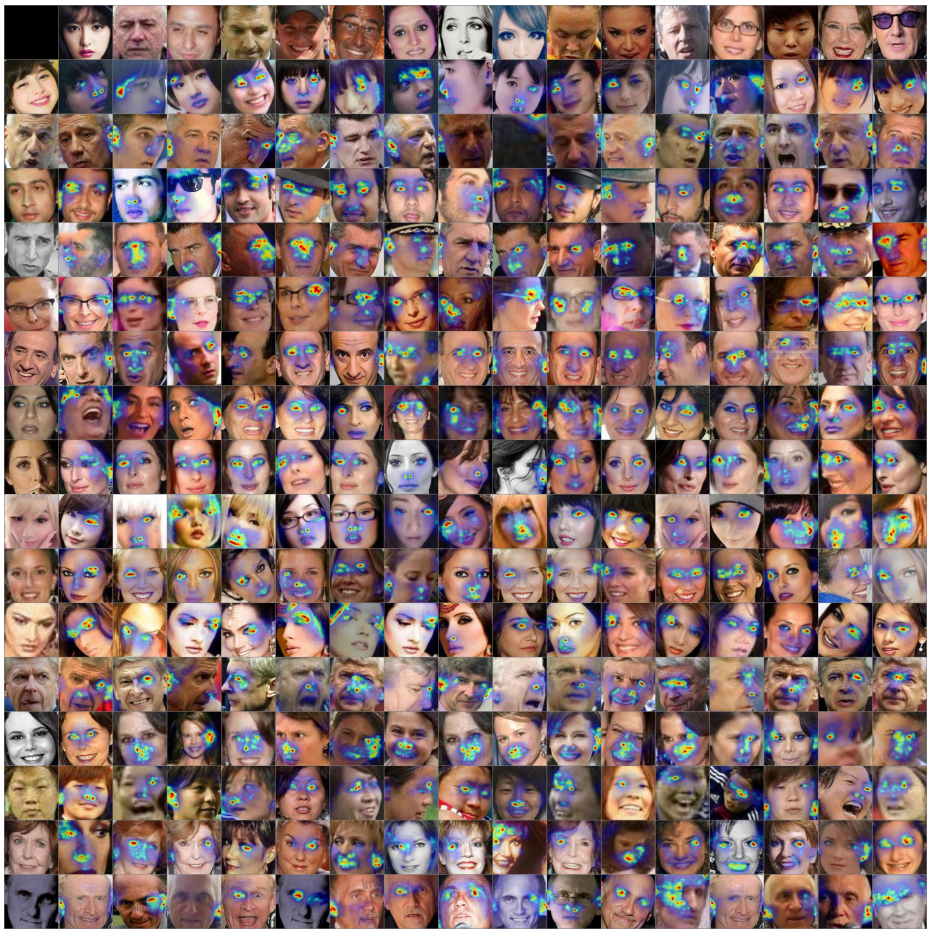

Fig. 4. Contrastive triplet EBP (ResNet-101). This montage shows the contrastive EBP for a ResNet-101 whitebox. Observe that this saliency map is unstable, and at times generates saliency maps on the background of the image (e.g. probe (16,15)). This is a known challenge of contrastive EBP, which led towards the development of truncated contrastive EBP.

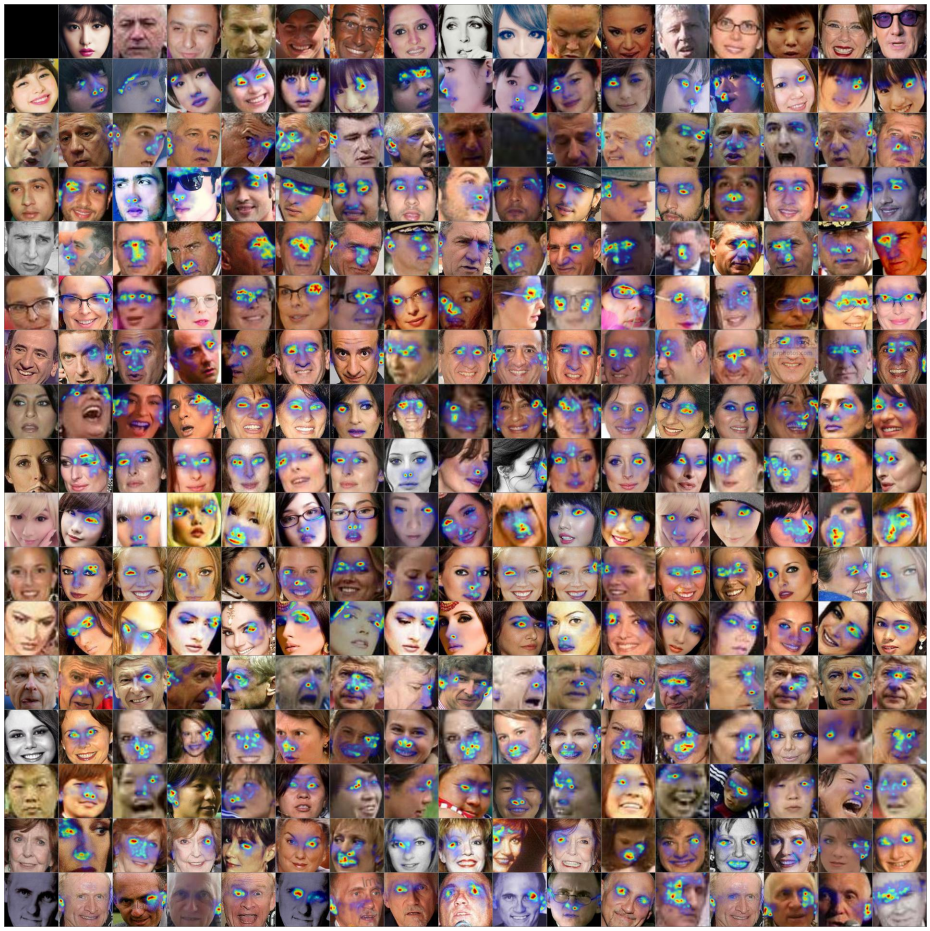

Fig. 5. Truncated contrastive triplet EBP (ResNet-101). This montage shows truncated contrastive triplet EBP.

Fig. 6. Subtree Triplet EBP (ResNet-101). This montage shows subtree triplet EBP.

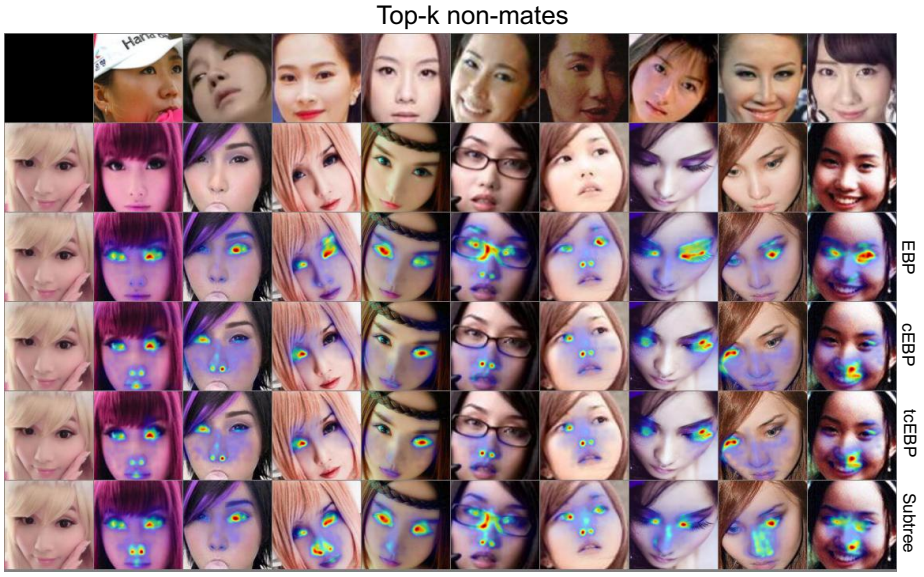

Fig. 7. Single probe montage (ResNet-101). This montage compares four white box methods on a common set of probes, such that the non-mates are now ordered in decreasing similarity with the mate. This shows how the different methods compare for real-world doppelgangers.

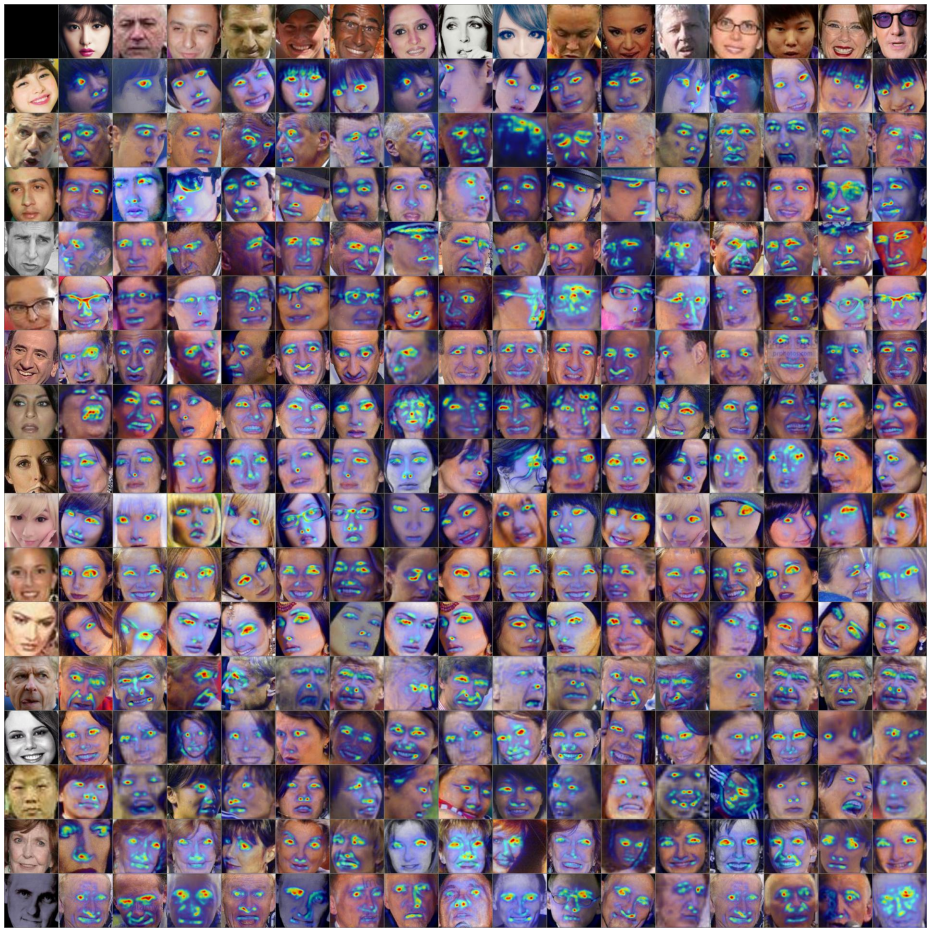

Fig. 8. EBP (Light-CNN [2]). This montage generates the EBP saliency map for the Light-CNN network. This should be compared with Fig. 3, which shows that this network exhibits more saliency around the mouth and brow than the ResNet-101 network.

Fig. 9. Contrastive triplet EBP (Light-CNN [2]). This montage should be compared with Fig. 4 to show the differences for contrastive triplet EBP comparing ResNet-101 with light-CNN.

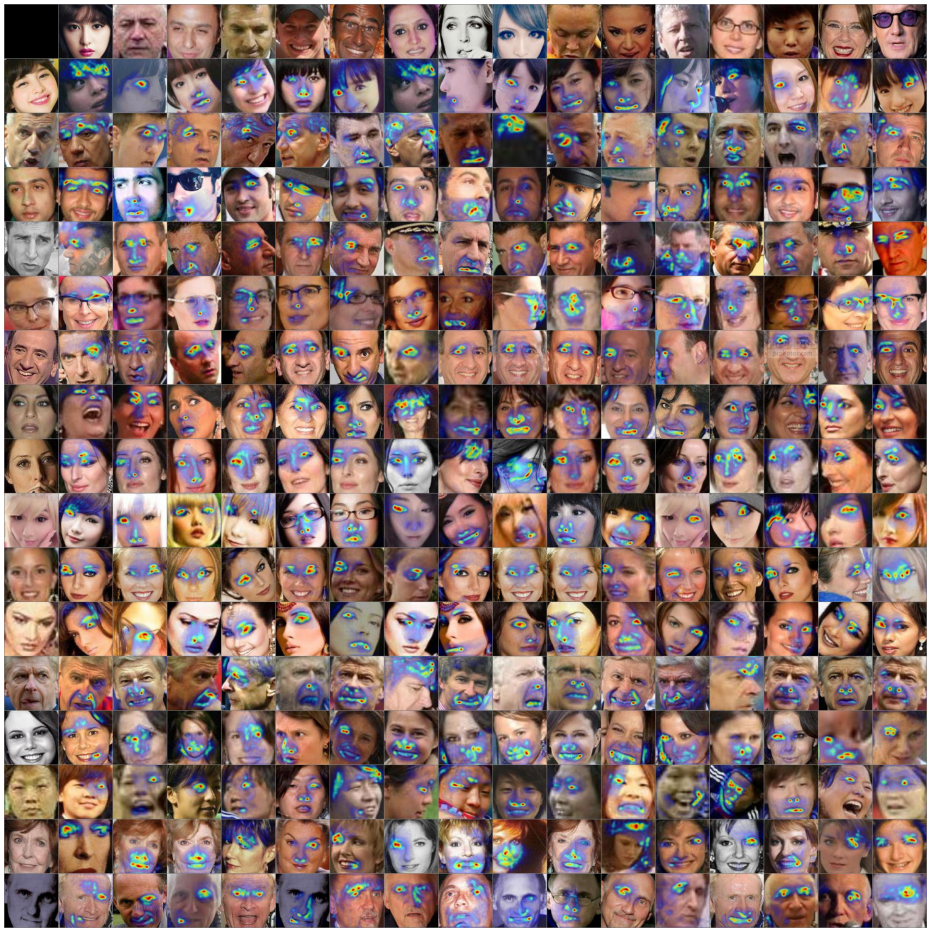

Fig. 10. Truncated contrastive triplet EBP (Light-CNN [2]). This montage should be compared with Fig. 5 to show the differences for tcEBP comparing ResNet-101 with light-CNN.

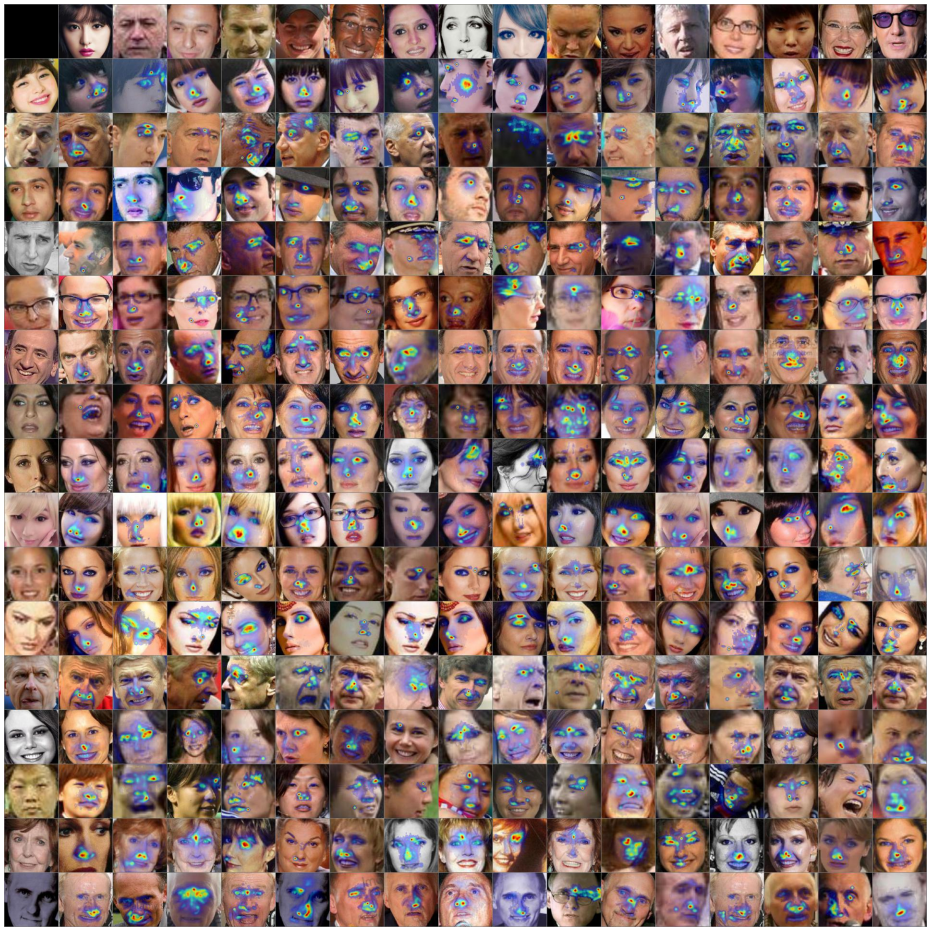

Fig. 11. Subtree triplet EBP (Light-CNN [2]). This montage should be compared with Fig. 6 to compare the differences for subtree triplet EBP for ResNet-101 vs. light-CNN.

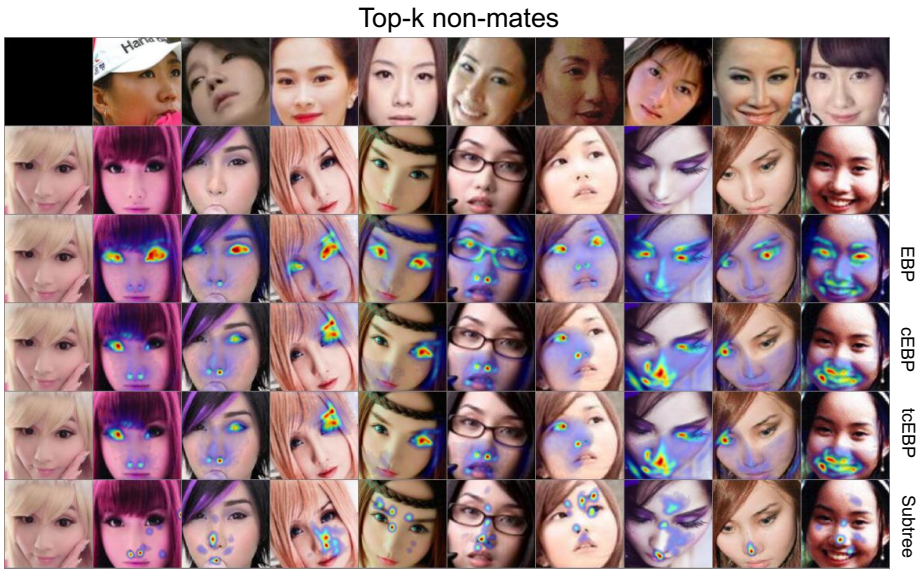

Fig. 12. Single probe montage (Light-CNN [2]). This montage should be compared with Fig. 7 to compare the effect of top-k non-mates for ResNet-101 vs. light-CNN.

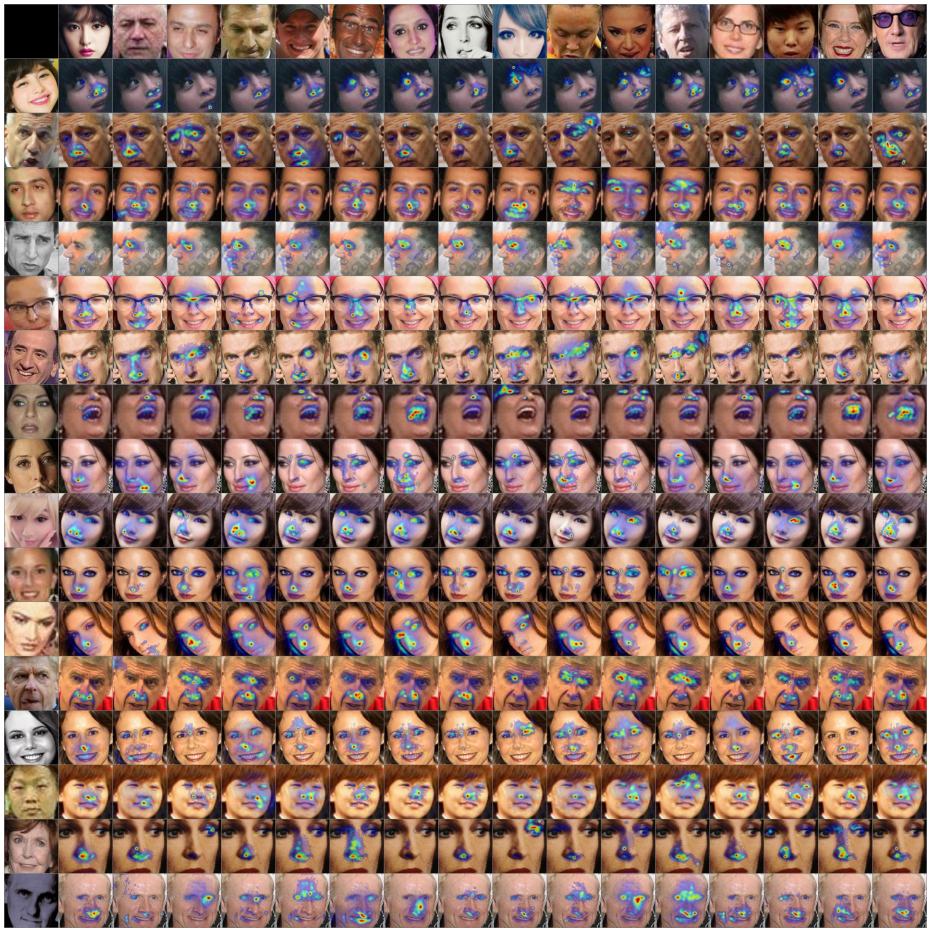

Fig. 13. Repeated probe montage (Light-CNN [2]). This montage shows the same probe repeated across each row to highlight the effect of the non-mate in the triplet on the resulting saliency map.

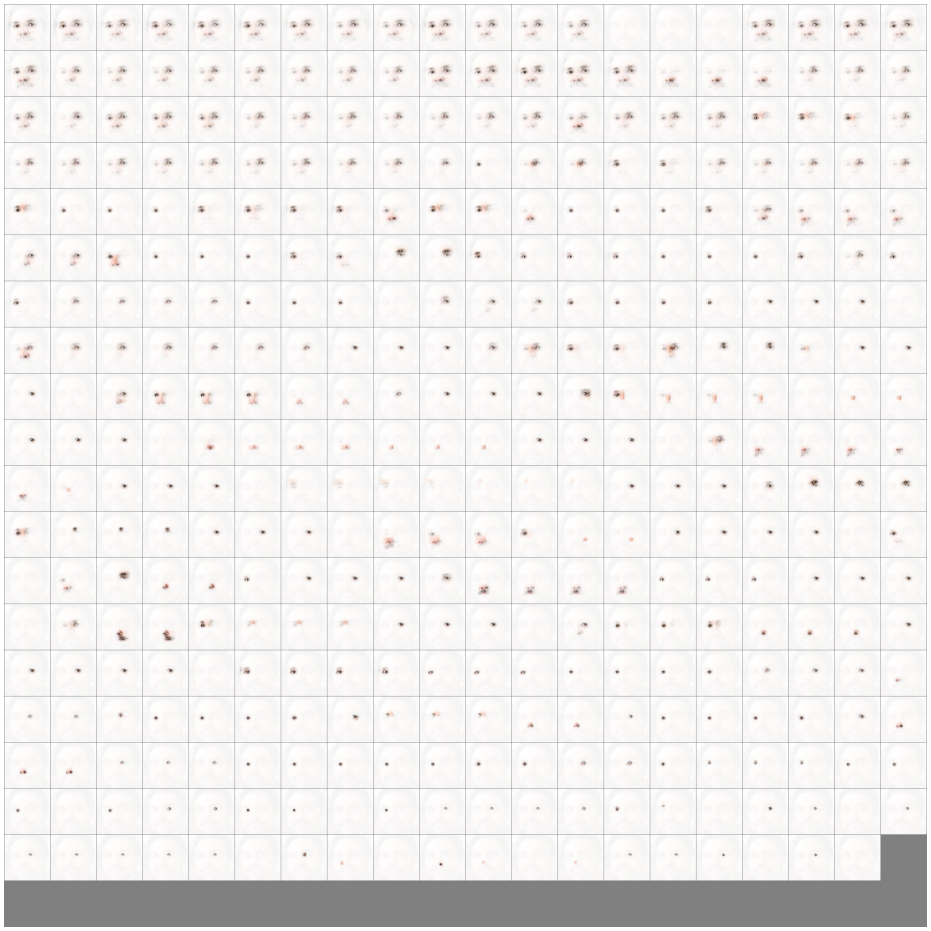

Fig. 14. Layerwise EBP. This montage shows the EBP saliency map generated starting from the maximum excitation for each layer in a ResNet-101 network. The layers are ordered rowwise, starting from the embedding layer in the upper left down to the image layer in the bottom right. The visualization shows the saliency map encoded as the alpha channel of a cropped face image, so that non-zero saliency results in a more opaque (less transparent) region. This visualization style is useful to accentuate small activations. This result shows that saliency maps starting from the layers closer to the embedding result in holistic regions covering the eyes and nose, layers in the middle show parts such as the eyes, nose and mouth, layers closer to the image are highly localized on specific regions of the image, and some layers provide no excitation at all.

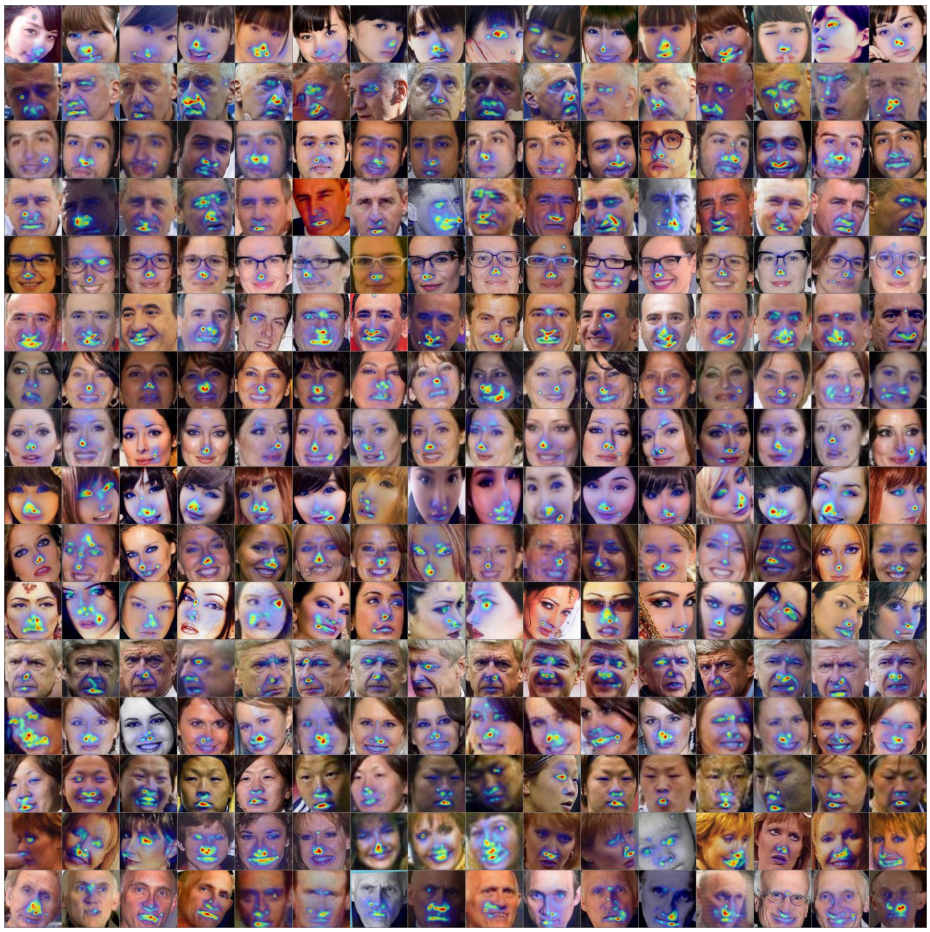

Fig. 15. Qualititative visualization study. This figure shows the XFR saliency maps generated using the LightCNN Subtree EBP method for 16 probes (columns) of 16 subjects (rows), each with 16 mates (not shown) and a common set of 8000 nonmates (not shown) all sampled from VGGFace2 [1]. Results show that the discriminative features used to distinguish a subject from the entire nonmate population are primarily the nose and mouth for frontal probes and eyes for non-frontal probes. These network attention maps are remarkably consistent across probes and provide insight into the features that a network uses to distinguish a subject from a large set of nonmates (i.e. What makes you unique?).

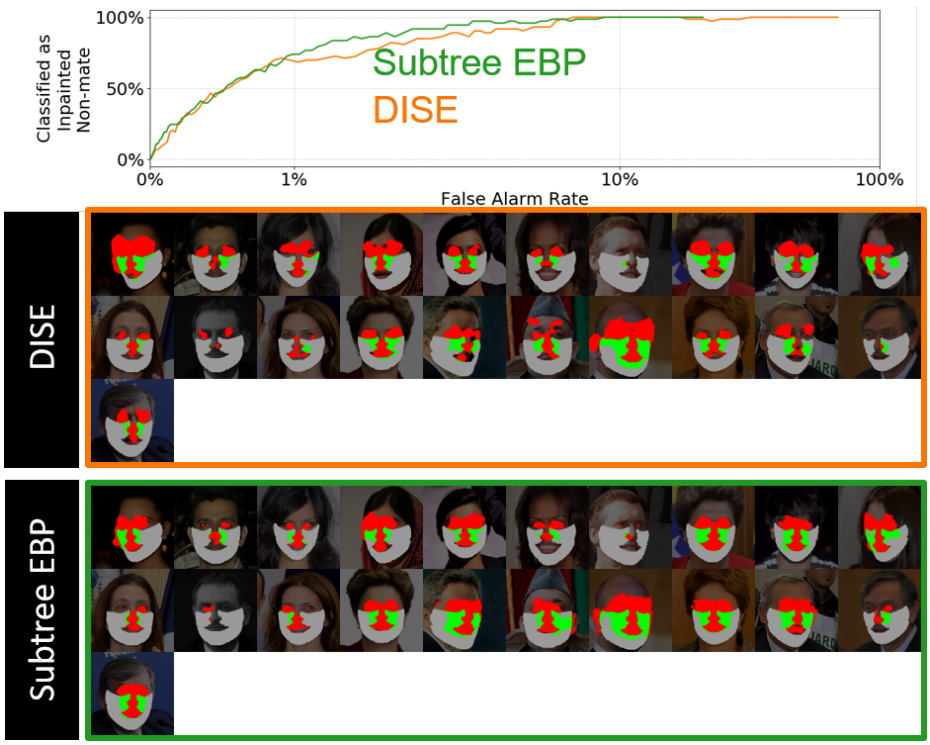

Fig. 16. Cheek/Chin Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.

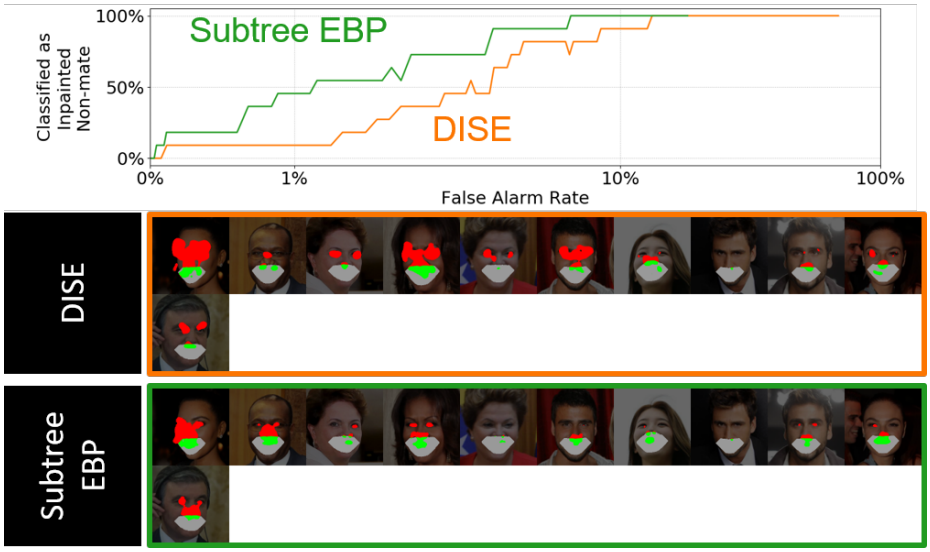

Fig. 17. Mouth Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.

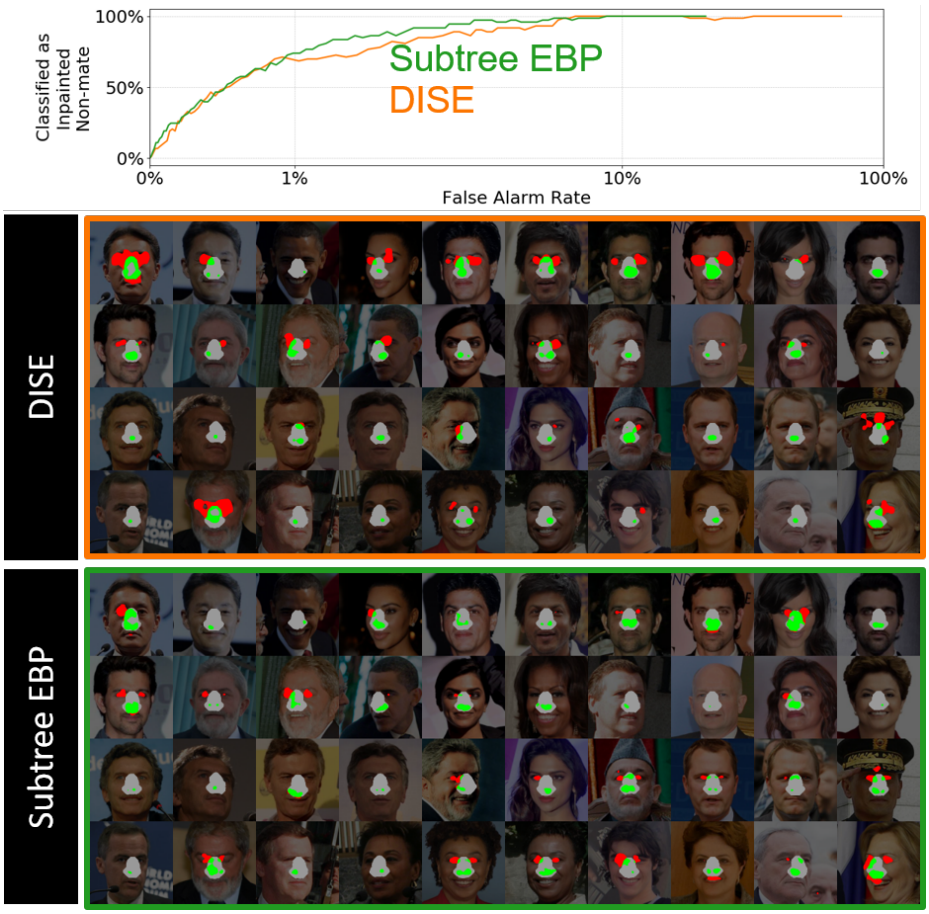

Fig. 18. Nose Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.

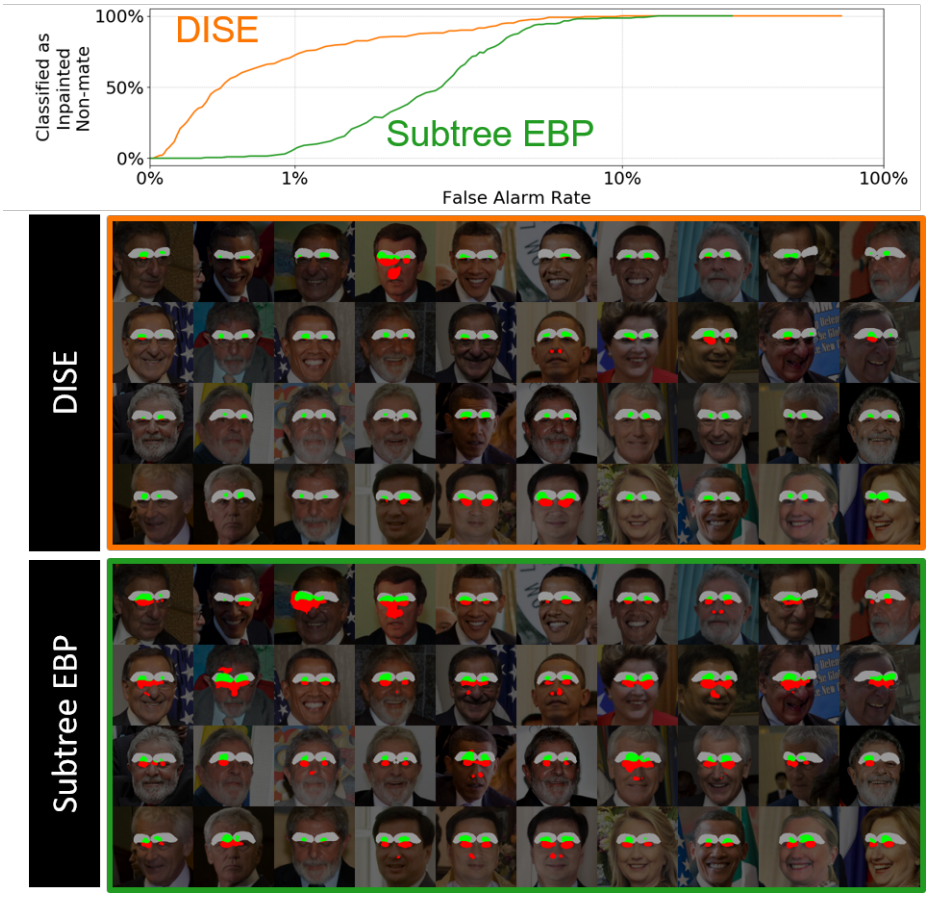

Fig. 19. Eyebrow Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.

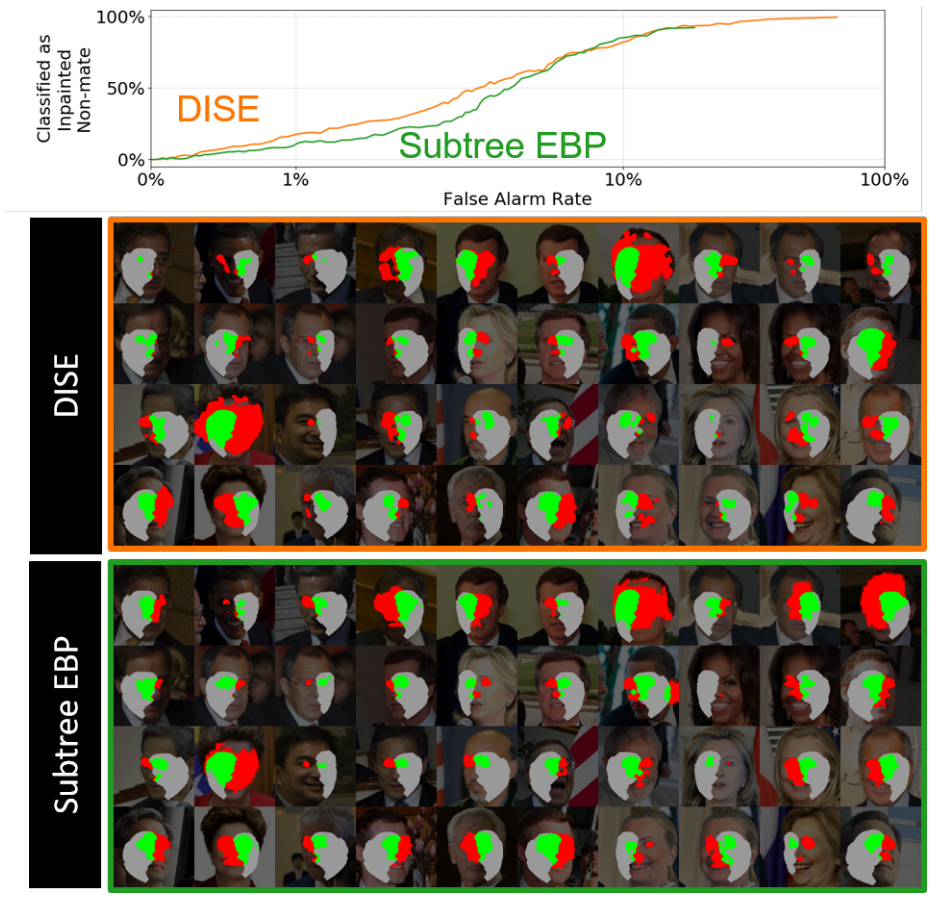

Fig. 20. Left-/Right-Face Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.

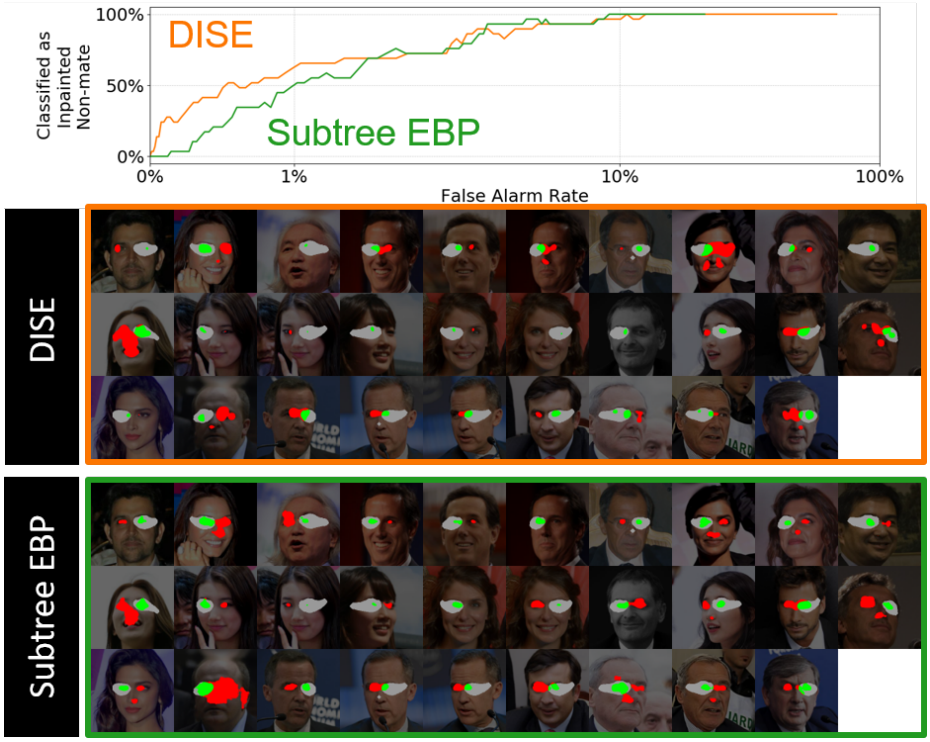

Fig. 21. Left-/right- eye Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.

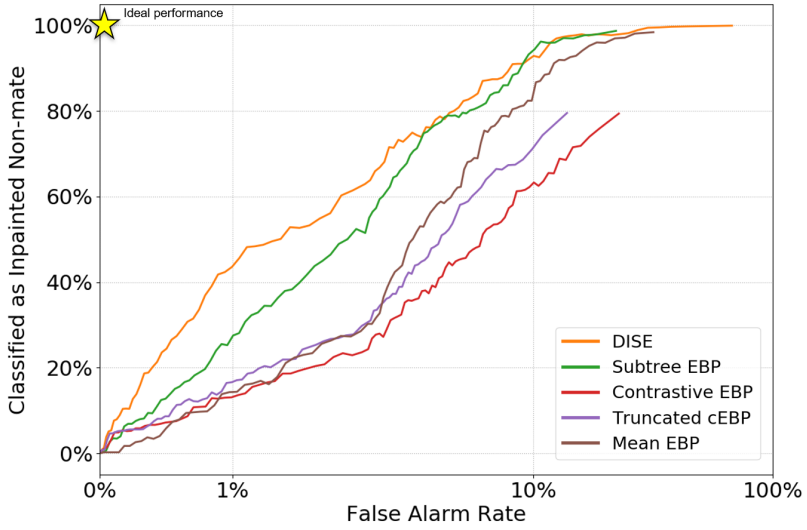

Fig. 22. Inpainting game analysis using the ResNet-101.